

# Cranberry microsatellite marker development from assembled next-generation genomic sequence

Laura Georgi · Roberto H. Herai · Ramon Vidal ·  
Marcelo Falsarella Carazzolle · Gonçalo Guimarães Pereira ·  
James Polashock · Nicholi Vorsa

Received: 18 April 2011 / Accepted: 18 July 2011  
© Springer Science+Business Media B.V. 2011 (outside the USA) 2011

**Abstract** The large-fruited cranberry (*Vaccinium macrocarpon* Ait.) is a native North American fruit that is a rich source of dietary phytochemicals with demonstrated and potential benefits for human health. Cranberry is a perennial, self-fertile  $2n = 2x = 24$  diploid, with a haploid genome size of about 570 Mbp. Present commercial cultivars are only a few breeding and selection cycles removed from their wild progenitors. With an irreducible minimum of 2 years per generation, and significant space and time requirements for phenotypic selection of traits of horticultural interest, genetic enhancement of cranberry could be

Persons wishing access to the cranberry sequence should contact the authors directly.

L. Georgi · N. Vorsa (✉)  
Marucci Center for Blueberry and Cranberry Research and Extension, Rutgers University, Chatsworth, NJ 08019, USA  
e-mail: Vorsa@aesop.rutgers.edu

L. Georgi  
e-mail: georgi@aesop.rutgers.edu

R. H. Herai · R. Vidal · M. F. Carazzolle · G. G. Pereira  
Laboratório de Genômica e Expressão, Instituto de Biologia, Universidade Estadual de Campinas, CP 6109, Campinas, SP 13083-970, Brazil

R. H. Herai  
UCSD Stem Cell Program, Department of Pediatrics, School of Medicine, University of California San Diego-UCSD, 9500 Gilman Drive-CMM-E, Room #2021D, La Jolla, CA 92093-0695, USA

facilitated by marker-assisted selection (MAS); however, the necessary resources, such as transcript or genomic sequences, molecular genetic markers, and genetic linkage maps, are not yet available. We have begun to generate these resources, starting with next-generation [sequencing by oligonucleotide ligation and detection (SOLiD) mate-paired] sequencing of an inbred cranberry clone, assembling the reads, and developing microsatellite markers from the assembled sequence. Evaluation of the resulting cranberry genomic microsatellite primers has provided a test of the accuracy of the sequence assembly and supplied much-needed molecular markers for a genetic linkage map of cranberry. Mapping these markers will permit sequence scaffolds to be anchored on the genetic map.

R. Vidal  
Laboratório Nacional de Biociências-CNPEM/ABTLuS, Campinas, Brazil

M. F. Carazzolle  
Centro Nacional de Processamento de Alto Desempenho em São Paulo, Universidade Estadual de Campinas, CP 6141, Campinas, SP 13083-970, Brazil

J. Polashock  
USDA-ARS, Chatsworth, NJ 08019, USA

**Keywords** Simple-sequence repeat (SSR) · Cranberry genome · Ericaceae

## Introduction

The large-fruited cranberry (*Vaccinium macrocarpon* Ait.) is a native North American fruit. Phylogenetically, it is positioned in the basal grade of the asterids, rather like the position of grape in the rosid clade (Angiosperm Phylogeny Group 2009). Like grape, it is a temperate woody perennial plant. Although asterids are as prominent a taxon as rosids and include some high-value crop species, they are underrepresented among sequenced plant genomes. The estimated dollar value of the US 2010 cranberry crop was US \$321 million (USDA NASS 2011). Beyond its dollar value, the fruit is valued as a “functional food,” and a rich dietary source of flavonoid antioxidants (Pappas and Schaich 2009). It has demonstrated antibacterial activity (Wu et al. 2008). Cranberry phytochemicals inhibit bacterial cell adhesion, and thus may protect against stomach ulcers (Neto et al. 2008) and tooth decay (Koo et al. 2010). Clinical trials have shown that cranberry consumption reduces risk of urinary tract infection (UTI) in young to middle-aged women (Guay 2009). There is evidence from a range of in vitro studies suggesting that consumption of cranberries has the potential to reduce the risk of cardiovascular disease (Ruel and Couillard 2007) and some forms of cancer, particularly of the gastrointestinal tract (Neto et al. 2008).

Cranberry is a self-fertile  $2n = 2x = 24$  diploid, with a haploid genome size of about 570 Mbp (Costich et al. 1993). Present commercial cultivars are only a few breeding and selection cycles removed from their wild progenitors. With an irreducible minimum of 2 years per generation, and significant space and time requirements for phenotypic selection of traits of horticultural interest, genetic enhancement of cranberry could be facilitated by marker-assisted selection (MAS); however, the necessary resources, such as transcript or genomic sequences, molecular genetic markers, and genetic linkage maps, are not yet available. In the present work, we describe our initial efforts to develop these resources, starting with SOLiD (Life Technologies, Carlsbad, CA) sequence analysis of an inbred cranberry clone and the

successful development of an initial set of microsatellite (simple sequence repeat, SSR) markers from the de novo assembled sequence. This type of marker has a number of desirable qualities, including high levels of polymorphism, codominance, reproducibility, transferability within and among related species, and abundance throughout the genome (Morgante and Olivieri 1993; Varshney et al. 2005, for example), and advances in DNA sequencing technology have facilitated identification of large numbers of microsatellite sequences for marker development. Microsatellite markers developed from our cranberry genomic sequence are being used to generate a genetic map of cranberry. The genetic map, in turn, will provide information about the positions in the genome of the sequence scaffolds that provided the microsatellite markers. This will lay a foundation for the genetic mapping of important traits such as biotic and abiotic stress resistance, yield, and fruit quality, and provide linked molecular markers for marker-assisted selection and the breeding of horticulturally superior cultivars.

## Materials and methods

### Plant material and DNA extraction

Young expanded leaves were collected from greenhouse-grown ramets of *Vaccinium macrocarpon* clone CNJ99-125-1, a fifth-generation inbred derived from self-pollination of the cultivar ‘Ben Lear’. This clone was chosen for sequencing because allelic variation is a major impediment to accurate de novo DNA sequence assembly. The initial level of heterozygosity theoretically should be halved by each generation of selfing, so CNJ99-125-1 ought to display less than 4% of the heterozygosity of its progenitor. Nuclear DNA extraction protocol A (Lutz et al. 2011), was modified as follows: Extraction buffers were freshly prepared from autoclaved stock solutions and molecular-biology-grade sucrose crystals and used the same day. One gram of leaf tissue was placed in each of two 35-ml Retsch (Haan, Germany) grinding jars with one 20-mm stainless-steel ball per jar and frozen in liquid nitrogen. The frozen tissue was disrupted on a Qiagen TissueLyser II (Hilden, Germany) for 1 min at 30 Hz, then the jars were placed again in liquid nitrogen and the

disruption was repeated. The powdered tissue was suspended in 200 ml cold extraction buffer 1 (0.4 M sucrose, 10 mM Tris-HCl pH 8, 10 mM MgCl<sub>2</sub>, 5 mM β-mercaptoethanol), filtered through two layers of Miracloth (Calbiochem, La Jolla, CA), and centrifuged at 2,000 × *g* for 20 min at 4°C. The pellet was washed three times in about 25 ml cold extraction buffer 2 (0.25 M sucrose, 10 mM Tris-HCl pH 8, 10 mM MgCl<sub>2</sub>, 5 mM β-mercaptoethanol, 1% Triton X-100) per wash and centrifuged at 12,000 × *g* for 10 min at 4°C. Pelleted nuclei were resuspended in 4 ml cold extraction buffer 3 (1.7 M sucrose, 10 mM Tris-HCl pH 8, 2 mM MgCl<sub>2</sub>, 5 mM β-mercaptoethanol, 0.15% Triton X-100) and pelleted through a cushion of extraction buffer 3 by centrifugation at 14,000 × *g* for 60 min at 4°C. The cleaned nuclei were lysed in 8 ml cetyltrimethylammonium bromide (CTAB) buffer [2% CTAB, 1.4 M NaCl, 100 mM Tris-HCl pH 8, 20 mM ethylenediamine tetraacetic acid (EDTA) pH 8, 100 mM β-mercaptoethanol] at 65°C for 15 min and extracted with chloroform. The crude nuclear DNA was digested with RNase A (50 µg/ml; Sigma-Aldrich, St. Louis, MO) for 30 min at 37°C, extracted again with chloroform, and precipitated with isopropanol. Isopropanol pellets were resuspended in water and reprecipitated with ethanol. The precipitated DNA was recovered by hooking with a sealed sterile Pasteur pipette, and further purified using a Qiagen Genomic-tip 100/G following the manufacturer's protocol.

Total DNA for polymerase chain reaction (PCR) testing of microsatellite primers was extracted from leaves (ca. 50 mg) of greenhouse-grown plants of cultivar 'Stevens' (a productive, widely adapted variety) and accession US88-70 [a variety with fruit rot resistance (Johnson-Cicalese et al. 2009)], using a modification of the method described by Stewart and Via (1993). Briefly, leaves were ground in CTAB buffer minus ascorbic acid and diethyldithiocarbamate, in 2-ml Safe-Lock (Eppendorf AG, Hamburg, Germany) tubes containing two 5-mm stainless-steel beads, on the TissueLyser II for 1.5 min at 30 Hz. Following incubation at 65°C and chloroform extraction, the crude DNA extract was digested for 30 min at 37°C with RNase A (20 µg) prior to isopropanol precipitation. DNA was resuspended in water, quantitated on a NanoDrop spectrophotometer (Wilmington, DE), and diluted to concentration of 5 ng/µl.

## Library construction and SOLiD sequencing

A 2 × 50 bp mate-paired library was constructed at the Waksman Genomics Core Facility (Piscataway, NJ) using 60 µg purified nuclear DNA from CNJ99-125-1. The DNA was sheared using a HydroShear apparatus (Digilab Inc., Holliston, MA), purified on a Qiaquick column (Qiagen, Germantown, MD), and end-repaired using an END-IT kit (Epicentre, Madison, WI). The fragmented, adaptor-ligated DNA was size-selected on a 1% agarose TAE (40 mM Tris acetate, 1 mM EDTA pH 8) gel; fragments in the 1.5–2-kbp range were recovered using Qiagen's Qiaquick gel purification protocol. These fragments were circularized by ligation to an internal biotinylated adaptor using Quick Ligase (New England Biolabs, Ipswich, MA) and used for library construction following the SOLiD 3 Plus 2 × 50 bp mate-paired library preparation protocol (Applied Biosystems, Carlsbad, CA). Following size selection on a 3% agarose TAE gel, the library (250–350-bp fragments) was purified using a Qiagen MinElute column and quantitated as directed in the library preparation protocol using a TaqMan (Applied Biosystems) assay prior to emulsion PCR and sequencing on the SOLiD 3 Plus apparatus.

## De novo sequence assembly

Assembly of the cranberry genome used an SGI Altix cluster based on Intel Itanium 2 processors with 158 central processing units (CPUs, 246 cores) and 7,750 Gb of random-access memory (RAM). The machines are connected by NUMAFlex generation 4 technology and InfiniBand connectivity, permitting a process to address 176 Gb of RAM. These systems are available at CENAPAD-SP (National Center for High-Performance Computing) in São Paulo, Brazil.

Following correction and exclusion of low-quality reads with the SOLiD Accuracy Enhancement Tool (SAET, <http://solidsoftwaretools.com/gf>) and CSFastaQualityFilter script from Applied Biosystems, reads were converted to double-encoded format using the script solid\_denovo\_preprocessor.pl for input into the Velvet assembler (Zerbino and Birney 2008). Single-read assemblies were performed using a wide range of different *k*-mers, and the optimal *k*-mer was found to be 41. Consequently, a paired-end assembly [insert size = 1,700, standard deviation (SD) = 500] was

run with this  $k$ -mer value. The resulting contigs were decoded to base space using the script solid\_denoovo\_postprocessor.pl followed by the deNovoAdp program. As deNovoAdp broke the scaffolds in the gap regions, an in-house script was used to rejoin contigs into scaffolds. All reads were decoded to base space to enable use of the GapCloser script from SOAPdenovo (Li et al. 2010) to fill in gaps in the decoded assembly with nucleotide sequences.

#### Microsatellite marker development

Sequences of the 46 largest scaffolds, plus two additional scaffolds, were submitted to the SSR tool on the Genome Database for Rosaceae (GDR) website (<http://www.rosaceae.org>, Jung et al. 2008) to identify microsatellites and generate primer sets for their amplification. The additional scaffolds were included because they contained sequences of interest: Scaffold 252 contained DNA sequence that potentially encodes a MADS box similar to *Prunus persica* dormancy-associated MADS box DAM1 (Bielenberg et al. 2008), although the similarity appears to be confined to the MADS-box domain. Scaffold 15903 contained portions of two putative UDP-glycosyltransferase genes. Primer pairs (Table 1) were synthesized by Integrated DNA Technologies (IDT, Coralville, IA) with an 18-bp M13 extension (5'-TGTAAAACGACGGCCAGT-3') on the 5' end of the forward primer in each pair, to permit labeling of fragments by PCR with a fluorescently tagged M13 primer (Oetting et al. 1995; Schuelke 2000); M13 primers tagged with WellRED D2, D3, and D4 (Beckman-Coulter, Fullerton, CA) were also obtained from IDT. Amplification reactions were performed in 10  $\mu$ l volumes containing 1  $\times$  Colorless GoTaq Flexi buffer (Promega Corporation, Madison, WI), 2 mM MgCl<sub>2</sub>, 0.2 mM dNTPs (each), WellRED-dye-labeled M13 primer and microsatellite reverse primers at 0.3  $\mu$ M each, 0.075  $\mu$ M microsatellite forward primer (with 5' M13 extension), 3 ng genomic or total DNA of cranberry clones CNJ99-125-1, US88-70, or 'Stevens', extracted as described above, and 0.25 units GoTaq Hot Start polymerase (Promega Corporation) per reaction. For fragment cloning and sequencing, reaction volumes were doubled, the forward and reverse microsatellite primers were supplied at equimolar concentrations (0.3  $\mu$ M each), and the labeled M13

primer was omitted, except for two primer sets (scf1h and scf3a) that only produced amplification product when the M13 primer was included. Applied Bio-systems (Life Technologies) thermal cyclers (GeneAmp PCR system 9700 or Veriti) were programmed as follows: For fluorescent labeling reactions, an initial 3-min denaturation step at 94°C was followed by 30 cycles of 40 s at 94°C, annealing at 52°C for 45 s, and extending at 72°C for 45 s, then an additional 8 cycles in which the annealing temperature was increased to 53°C, and ending with a 30-min incubation at 72°C. Fluorescent fragment analysis was performed on a CEQ 8000 genetic analysis system (Beckman-Coulter) using the DNA size standard kit-600 (Beckman-Coulter) and the machine's Frag-4 separation method. For nonfluorescent reactions, an initial 2-min denaturation step at 95°C was followed by 30 cycles of 95°C for 30 s, annealing at 52°C for 30 s, extending at 72°C for 30 s, and ending with a 5-min incubation at 72°C.

PCR products were cloned using the pGEM-T Vector system (Promega) and transformed into chemically competent *Escherichia coli* DH5 $\alpha$  (Life Technologies, Carlsbad, CA). Two colonies from each transformation were grown up in Luria broth (LB) for plasmid DNA purification using a Zippy plasmid miniprep kit (Zymo Research, Irvine, CA) and sequenced in both directions on the CEQ 8000 using the GenomeLab DTCS-quick start kit (Beckman-Coulter), following the manufacturer's recommendations for 10  $\mu$ l dye terminator cycle sequencing reactions. Sequencher 4.10.1 (Gene Codes Corporation, Ann Arbor, MI) was used to assemble the plasmid sequences.

## Results and discussion

### Sequence assembly

SOLiD sequence reads of the inbred cranberry were obtained totaling 32 Gbp, for an approximately 58-fold coverage of the genome. The final Velvet assembly (Table 2) contained 441,159 contigs in 68,496 scaffolds larger than 300 bp, for a total length of 566.7 Mbp, with 258 Mbp in gap regions and a scaffold N50 of 26,335 bp. The largest scaffold was 288,666 bp. Despite the use of only short reads in the assembly process of the cranberry genome, the

**Table 1** Summary of cranberry genomic microsatellite markers derived from assembled SOLiD mate-paired sequence reads

Marker ID	Primer sequences (5'-3') <sup>a</sup>	Repeat motif	Predicted no. of repeats	Predicted product (bp) <sup>b</sup>	Amplified product(s) (bp) <sup>c</sup>	Polymorphic <sup>d</sup>
scf1h	CCGTGGAGGAGAATGGTTA TTCCGATGCACAAGATATGG	tta	12	281	na <sup>e,f</sup>	No
scf2s	TGAGACGTACGCACTAGCCA GTCGATGGTGTTCGATG	ct	21	207	165	Yes
scf3a	CGTTCTAACAGAGCAACTGCACG AACGGCACGATTCTGTTTAC	tc	19	144	na <sup>f</sup>	No
scf4b	GATACGATACGGATAACGCGG GTCGATCATGGTCGTCAGTG	ga	15	266	310	Yes
scf5k	GCATTACTAACAGCATCCAA GAGCCACTTTCACTCCAA	tc	20	262	248	Yes
scf6q	ACCACCAAGAACACATCAA AATGGAGGAGTGTCACCTG	ga	18	162	na	No
scf7n	TGCCGTGTTGGATGACTAA AATGAAAATAGCCATTGCGG	att	11	292	na	No
scf8l	CGAATCCGAAGATCAGAAC GGGATACCAAGAGATTCCCG	ag	20	172	157	Yes
scf9x	TCATGCGTCGATTTCAGAAC GCATGAAGCTTGTCAAGACACC	tg	22	212	na	No
scf10k	AAGGAACCGATCGAGGAAC TCACATTCTCGTGTGAGGC	ag	11	127	127	No
scf11i	TCTCTTATGGCCTAACCGA CCACGCCACAATATTTCTT	ag	15	220	173	No
scf12i	GACCGTAAGCGTGGATTGTT TCCTACCACTACCACCACTGC	ag	16	244	207	Yes
scf13a	TAGAGGGCGTTGAAAGGAGA CCCCAAATTCTCCCCATTA	ga	17	300	319	Yes
scf14j	CAGCAGAACATTCAAGGAAAGCC AGCTTTCCACACGCTCATT	ag	14	170	198	No
scf15a	ATCTCCCACCTACCCCAAAG GCATATCGACAATTCAAACCC	ga	10	274	222	No
scf16i	AGTTGCAAGGTCTGCTCCAT TTTCGATTACCGAACATTGCC	ag	18	235	239	Yes
scf17k	TCAGCGCGTCTGACAAGTAG TGGGAACGTATCGGCTAAAG	ag	20	206	na	No
scf18e	TGAGAACCAATTGGCAACA TGGAACGTTAAAAGGATGGG	ttg	11	223	na	No
scf19x	GGGTGAAATCTCGGCATTA AAGGTCCCTTCACATGTTGC	ga	18	190	184	No
scf20o	GTACGAAACCCACCTCCAGA TGACACCAAGAAAACACCCA	ag	18	176	x <sup>g</sup>	No
scf21g	AAGTCAGGGTACCAACACGC TGTAACTCGTTCGCAGGTG	ag	21	162	na	No
scf22m	TAACTTCACTAGCCCCACCCG AGGGTTTAGGCACCTAGGACA	ct	19	293	423	Yes

**Table 1** continued

Marker ID	Primer sequences (5'-3') <sup>a</sup>	Repeat motif	Predicted no. of repeats	Predicted product (bp) <sup>b</sup>	Amplified product(s) (bp) <sup>c</sup>	Polymorphic <sup>d</sup>
scf23d	TAGCTGTCCCCACTGGAATC CACATGGTATCAGAACCGGA	ag	19	292	na	No
scf24k	ATTGAGCCCCACACTACAGG AGCCATGGAAATCCAACAAA	ga	17	247	277	Yes
scf25m	GGTTAACAGCAACGCCCTTC CACCAAGGGAGTAGAAACGGA	ct	20	186	207	Yes
scf26r	ATGATGTTGGATGTGCCTCA TTCCTCAACAAACCCCTCCAC	ct	20	185	260	Yes
scf27l	GATTCAAGGCCAAGAATTCCA CACACACAGGACAAAGCCAC	ag	12	290	261	No
scf28b	GGTCAGTGTGTTGAGAGC GGTCCTGTACTACGCCCTTGC	ct	11	226	378	No
scf29j	TTACTCTCGCGTTGTGATGC CCTTTGTTGCATCCTCATTG	ag	16	183	na	No
scf30g	ATTGGAGCCCTAACACCAGG TCCGTATGCAAGTCCACAAC	ac	18	273	213, 215	Yes
scf31d	GCATATGAATGCCAACACAA TGATTGCAATTGGTCCCT	ag	19	217	179	No
scf32c	AACACAGAGTCCCCACTTGC TGAGGCTCTGTTCCAACTT	tc	21	190	na	No
scf33v	CCCTCTGCCAACACGTATT GGGGCTGAAGTCCACATTAA	tg	21	232	na	No
scf34s	TACCCGGCCGTATATGTAGC AATGTGACGTCAGAGGGAGG	ct	20	202	179	Yes
scf35f	TCCAAGTTAGTCTCGCGGT TGTCCGAATGGGTGTGTATG	ga	20	146	na	No
scf36l	AGTCCGTAAAGAGACATGCAG TTTGGGATCAAATCTCTCGG	ag	20	232	230	Yes
scf37h	TGGACTTTCTGCTTGGCT GGATACACGTGACCGAGCTT	ga	17	153	368	Yes
scf38b	ACTCCATCACCACACCGAA ACCCCTAACCAACCGTCTTC	ag	17	209	202	No
scf39e	GCGGAATAAGATCCCGTGA CCACACAAACCTGCTGCATAC	tc	21	219	202	Yes
scf40o	TGGTATGGTCAAAGCACA TTCTTCACGCTACTGCTGGA	ag	17	247	na	No
scf41c	GGTCCCGAAAAACACTCTGA ACGTCAGTCCATGCATTCAA	ag	10	243	250	No
scf42k	GGAAACCAAGTGGCAGAACAT ATTGGACATCAGAACACGCA	ag	16	250	188	No
scf43g	ATGGGCTCCATTGTGTTTG ATCGCCCCCTACCTCGTATCT	tc	18	206	171	Yes
scf44a	ACAAAACCACTGGCGAAAAC GAGTGACCAGGGAGATGAA	ag	19	249	259	Yes

**Table 1** continued

Marker ID	Primer sequences (5'-3') <sup>a</sup>	Repeat motif	Predicted no. of repeats	Predicted product (bp) <sup>b</sup>	Amplified product(s) (bp) <sup>c</sup>	Polymorphic <sup>d</sup>
scf45d	TTCTTGTGGTTGTGCTGCAT TAATGGCTGAAACGCTCACAA	ct	14	288	219	Yes
scf46g	AAAGGGAGCAATCTAACCA CAGCCAAACAGCTGATGATG	ga	21	210	204	Yes
scf252g	TTTCAATGCTTGCTTGG CTAACTAGGACCGGGCTTC	ag	17	165	na	No
scf15903c	ACTTACCCACGAGCCTACCA GAAGGAGAAAGTGACGTCGG	ct	22	299	294, 316	Yes

<sup>a</sup> Forward primers were synthesized with a 5' M13 primer tag: 5'-TGTAAAACGACGGCCAGT-3'

<sup>b</sup> Size of amplification product (in base pairs) predicted from sequence assembly

<sup>c</sup> Size (in base pairs) of product amplified from cranberry clone CNJ99-125-1, not including the M13 primer tag

<sup>d</sup> Segregating in an F1 cross between US88-70 (accession with fruit rot resistance) and 'Stevens' (widely adapted, productive cultivar)

<sup>e</sup> na no amplification

<sup>f</sup> Off-target amplification only in reactions containing the M13 primer

<sup>g</sup> No fluorescently labeled product detected, but an approximately 400-bp fragment was visible on ethidium-bromide-stained agarose gel

**Table 2** Results of de novo assembly of next-generation short-read sequences of cranberry nuclear genomic DNA

Estimated genome size	570 Mbp
Assembly length	566.7 Mbp
Number of scaffolds >300 bp	68,496
N50 scaffold number	6,023
N50 scaffold length	26,335 bp
Longest scaffold	288,666 bp
Average scaffold length	8,274.28
Bases in gaps	258 Mbp

assembly statistics showed that SOLiD color space reads can be used to perform low-cost assembly of plant genomes. At ca. 570 Mbp, the cranberry genome is the largest to be assembled to date using entirely SOLiD short-read sequences.

#### Microsatellite markers

Of the 48 primer pairs synthesized, 32 produced PCR amplification products. Two additional primer sets (scf1h and scf3a) only amplified in reactions containing the M13 primer. When these products were cloned and sequenced, they proved to be off-target amplifications flanked on one side by the M13 primer and the

other with the microsatellite reverse primer sequence (results not shown). An additional primer set (scf20o) yielded a product that could readily be visualized by ethidium bromide staining when electrophoresed on agarose gels, but was invisible to the Beckman-Coulter CEQ 8000. The most likely explanation is that this, too, is an off-target product lacking the (M13-tagged) forward microsatellite primer sequence.

No attempt was made to optimize reaction conditions, and some of the failed primer sets might conceivably have performed if, for example, the annealing temperature were lowered; however, the annealing temperature used was one that ought to have been suitable based on calculated melting temperatures for these primers. Other possible explanations for failed amplifications include base errors in the regions used to design the primer sequences or errors of assembly such that the primer sequences are either too distant (or unlinked) or not in the proper orientation to permit amplification or detection of the products. Sequence assembly errors are a likely explanation, given that most amplification products obtained differed to some extent from their predicted sizes.

The discrepancy between predicted and observed amplified fragment sizes was further investigated by cloning and (Sanger) sequencing products of amplification reactions using CNJ99-125-1 DNA as

**Fig. 1** Sanger sequences of microsatellite fragments amplified from cranberry clone CNJ99-125-1 (Query, upper sequence) aligned with the corresponding region of the cranberry SOLiD assembly (Sbjct, lower sequence). The targeted simple sequence repeat is in *uppercase*, while flanking sequences are *lower case*. Sequence gaps are represented by *dashes*, and *vertical lines* mark sequence identities. Numbers refer to the position of the sequences in their assemblies. **a** scf15a versus scaffold 15; *arrows* and *italics* indicate a 51-bp duplication present in the assembly but not the amplified fragment. **b** scf22m versus scaffold 22. **c** scf24k versus scaffold 24. **d** scf28b versus scaffold 28. **e** scf37h versus scaffold 37. **f** scf41c versus scaffold 41. These Sanger sequences have been deposited in GenBank under accession numbers JN230514–JN230519.

**Fig. 1** continued

the template with the following primer sets: scf15a, scf22m, scf24k, scf28b, scf37h, and scf41c. The lengths of the resulting consensus sequences (223, 421, 277, 377, 366, and 249, respectively) were in good agreement ( $\pm 2$  bp) with the sizes estimated by fluorescent capillary electrophoresis (Table 1). The duplicate reads were virtually identical, differing by no more than one single base (scf22m, scf24k) and/or a single copy of a mono- or dinucleotide repeat (scf22m, scf41c). At this depth of sequencing, we cannot say whether these differences represent allelic variation or PCR artifacts. All six sequences aligned to their expected scaffolds in the cranberry genome assembly (Fig. 1). The assembly is composed of SOLiD reads with a maximum length of 50 bp. Reads containing simple sequence repeats are difficult to position in an assembly, because the repeated sequence proper occurs in multiple genomic locations. As the length of the repeat increases, it becomes less likely that a single 50-bp read will span it and contain sufficient flanking unique DNA sequence to position the repeat. Mate-pair information is used to assign repeats to their proper location in the assembly, based on their linkage to unique sequences and the known range of sizes of the sequenced fragments, which in this case varied from 1.5 to 2 kbp. Given the technical difficulty of assembling repeats, it is not surprising that the observed fragment sizes diverged from the predicted sizes. In five of the six cases examined, the microsatellite itself was where the alignments broke down.

The assembly contained a known gap for scf22m and scf41c. Given the variation in size of the fragments used to make the mate-pair library, it is not surprising that the point estimates for the lengths of the gaps deviated from what was observed. For scf24k, scf28b, and scf37h, the assembly juxtaposed sequences that should have been separated by gaps. In addition, the length of the microsatellite itself was underestimated in scaffolds 24 and 41. On the other hand, scf15a and scaffold 15 sequences aligned perfectly around the microsatellite, but the assembled sequence had two copies in tandem of a 51-bp sequence that was present only once in the scf15a cloned fragment. The absence of one copy of the 51-bp sequence from the cloned fragment does not appear to be a PCR artifact, as there is no sign of the larger fragment in the capillary electrophoresis trace (data not shown). Because the SOLiD and Sanger

sequences used different extractions of DNA from the cranberry clone, we cannot exclude the possibility that a deletion occurred in the plant. Illumina sequence data currently in assembly may help determine which sequence—the assembled SOLiD or the Sanger—is correct for this region.

Successful PCR-based marker development requires not only successful amplification, but also amplification of polymorphic products. Given the shortness of the SOLiD reads and the notorious difficulty of assembling repeated sequences (Miller et al. 2010), we were pleasantly surprised at the rate of successful amplification (67%, not including the three problematic primer sets) using the 48 microsatellite primer pairs designed from the assembled cranberry genomic sequence. Of these, 21 amplified polymorphic products that segregate in the progeny of a cross (US88-70  $\times$  'Stevens'), in which fruit rot resistance is also segregating. Before the cranberry inbred was sequenced, the best source of microsatellite markers for use in cranberry was heterologous blueberry (*V. corymbosum*) microsatellite markers. Of 39 blueberry expressed sequence tags (EST) and 10 blueberry genomic microsatellite primer sets evaluated on 7 cranberry accessions, 32 EST and 6 genomic sets supported amplification, with 18 EST and 5 genomic sets revealing polymorphism (Bassil et al. 2009). Twelve of these (seven EST and five genomic) are segregating in the US88-70  $\times$  'Stevens' progeny. The blueberry primers had previously been demonstrated to amplify polymorphic loci in blueberry (Boches et al. 2005), which might be expected to improve their success rate in cranberry. Nonetheless, the previously untested cranberry primers yielded more mappable markers (21 out of 48 primer sets) than did the blueberry primers (12 out of 49).

## Conclusions

We have begun to develop genomic resources for the large-fruited American cranberry, starting with the SOLiD sequencing of a mate-paired library using nuclear DNA extracted from an inbred clone of cranberry. Assembling the sequencing reads produced scaffolds with a total length, approximating the expected size of the cranberry genome, with more than half of that in contigs, albeit with numerous

gaps. The successful development of microsatellite primers from this assembly is a testament to its essential accuracy, even in regions (simple sequence repeats) that are a challenge to assemble. Many of these microsatellites are segregating in a sibship in which resistance to fruit rot is also segregating. Adding these much-needed markers to our nascent first-generation genetic map of cranberry will expand the map while at the same time providing presumptive genetic locations for the sequence scaffolds from which the markers were developed. By providing molecular markers linked to fruit rot resistance and other traits of interest, these genomic resources will assist in the selection of desirable genotypes for breeding of superior cultivars of cranberry.

**Acknowledgments** Funding for this work was provided by USDA SCRI grant number 2008-51180-04878, with additional funding from Ocean Spray Cranberries, Incorporated. We thank Dayani Stinson for technical support and Mark Diamond for assistance in manuscript preparation.

## References

- Angiosperm Phylogeny Group (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot J Linn Soc* 161:105–121
- Bassil N, Oda A, Hummer KE (2009) Blueberry microsatellite markers identify cranberry cultivars. *Acta Hortic* 810: 181–186
- Bielenberg DG, Wang YE, Li Z, Zhebentyayeva T, Fan S, Reighard GL, Scorza R, Abbott AG (2008) Sequencing and annotation of the evergrowing locus in peach [*Prunus persica* (L.) Batsch] reveals a cluster of six MADS-box transcription factors as candidate genes for regulation of terminal bud formation. *Tree Genet Genomes* 4:495–507
- Boches PS, Bassil NV, Rowland LJ (2005) Microsatellite markers for *Vaccinium* from EST and genomic libraries. *Mol Ecol Notes* 5:657–660
- Costich DE, Ortiz R, Meagher TR, Bruederle LP, Vorsa N (1993) Determination of ploidy level and nuclear DNA content in blueberry by flow cytometry. *Theor Appl Genet* 86:1001–1006
- Guay DRP (2009) Cranberry and urinary tract infections. *Drugs* 69:775–807
- Johnson-Cicalese J, Vorsa N, Polashock J (2009) Breeding for fruit rot resistance in *Vaccinium macrocarpon*. *Acta Hortic* 810:191–198
- Jung S, Staton M, Lee T, Blenda A, Svancara R, Abbott A, Main D (2008) GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. *Nucleic Acids Res* 36 (Database issue): D1034–D1040
- Koo H, Duarte S, Murata RM, Scott-Anne K, Gregoire S, Watson GE, Singh AP, Vorsa N (2010) Influence of cranberry proanthocyanidins on formation of biofilms by *Streptococcus mutans* on saliva-coated apatitic surface and on dental caries development in vivo. *Caries Res* 44:116–126
- Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Gao S, Kristiansen K, Li S, Yang H, Wang J, Wang J (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20: 265–272
- Lutz KA, Wenqin W, Zdepski A, Michael TP (2011) Isolation and analysis of high quality nuclear DNA with reduced organellar DNA for plant genome sequencing and resequencing. *BMC Biotechnol* (accepted for publication)
- Miller J, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. *Genomics* 95:315–327
- Morgante M, Olivieri AM (1993) PCR-amplified microsatellites as markers in plant genetics. *Plant J* 3:175–182
- Neto CC, Amoroso JW, Liberty AM (2008) Anticancer activities of cranberry phytochemicals: an update. *Mol Nutr Food Res* 52:S18–S27
- Oetting WS, Lee HK, Flanders DJ, Wiesner GL, Sellers TA, King RA (1995) Linkage analysis with multiplexed short tandem repeat polymorphisms using infrared fluorescence and M13 tailed primers. *Genomics* 30:450–458
- Pappas E, Schäich KM (2009) Phytochemicals of cranberries and cranberry products: Characterization, potential health effects, and processing stability. *Crit Rev Food Sci Nutr* 49:741–781
- Ruel G, Couillard C (2007) Evidences of the cardioprotective potential of fruits: the case of cranberries. *Mol Nutr Food Res* 51:692–701
- Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. *Nat Biotechnol* 18:233–234
- Stewart CN, Via LE (1993) A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR applications. *BioTechniques* 14:748–750
- USDA NASS (2011) Noncitrus fruits and nuts 2010 Preliminary Summary. <http://usda.mannlib.cornell.edu/usda/nass/NoncFruNu//2010s/2011/NoncFruNu-01-21-2011.pdf>. Accessed 18 April 2011
- Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. *Trends Biotechnol* 23:48–55
- Wu VCH, Qiu X, Bushway A, Harper L (2008) Antibacterial effects of American cranberry (*Vaccinium macrocarpon*) concentrate on foodborne pathogens. *LWT-Food Sci Technol* 41:1834–1841
- Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821–829